Hand-out for Discussion of EIA-826 Data Analysis
By Nancy Kirkendall and Joe Sedransk

Let , $\hat{e}_{ks} = y_{ks} - \hat{y}_{ks}$ represent the error in using the beta based on the sampled companies data to estimate for the nonsampled company, k, in region, s. Let n_s represent the number of sampled companies in stratum s; and N_s represent the total number of companies within the population in stratum s.

1) The average estimation error within stratum, s, is given by $\bar{e}_s = \frac{1}{N_s - n_s} \sum_{k=1}^{N_s - n_s} \hat{e}_{ks}$.
This is expressed as a percent by dividing by the total of the nonsampled companies in stratum s, $Tns_s = \sum_{k=1}^{N_s - n_s} y_{ks}$.

2) The RMSE of the estimation errors in stratum s is computed as
$RMSE_s = sqrt(\frac{1}{N_s - n_s} \sum_{k=1}^{N_s - n_s} \hat{e}_{ks}^{\,2})$.
This is expressed as a percent by dividing by Tns_s

3) The CV is the RMSE divided by average of nonsampled units $Tns_s / (N_s - n_s)$.
4) Percent error in estimating the nonsampled companies in the stratum:
$E_s = (N_s - n_s)\bar{e}_s * 100 / Tns_s$.
5) And the percent error in estimating the stratum total.
$E_s = (N_s - n_s)\bar{e}_s * 100 / (Tns_s + Ts_s)$. Where $Ts_s = \sum_{k=N_s - n_s + 1}^{N_{ss}} y_{ks}$

These summary statistics are also computed at the US level, and these are displayed for residential sales, commercial sales, and industrial sales in the attached tables. There may still be some anomalies in the data.

The table in each attachment shows the baseline case with the current stratification and data from IOUs not used in estimating for the nonsampled companies. There are 6 alternatives shown in the row of that top table.

a. Beta=1 (ie use last year's data as an estimate for this year's)
b. Estimation with gamma=.5
c. Estimation with gamma=.8
d. Estimation with gamma=.5 and one pass of outlier detection and removal
e. Estimation with gamma=.8 and one pass of outlier detection and removal
f. Estimation of gamma using a two stage least squares procedure
g. Estimation of gamma using a two stage least squares procedure and one pass of outlier detection and removal.

The second table in each attachment shows the results using current stratification but including the IOUs among the sampled companies in estimation for the nonsampled companies. The six alternatives above are repeated.

The third table uses an alternative stratification of the States into estimation groups as discussed below, and does not include IOUs in the estimation for non-sampled companies. The six alternatives above are repeated.

Finally the fourth table uses the alternative stratification and includes data from the IOUs among the sampled companies to estimate for the nonsampled companies.

Nancy's conclusions from reviewing the tables.

1. Gamma=.5 is better than gamma=.8. Using two stage least squares is very promising.
2. Using an automated approach to removing outliers and influential observations improves the estimation for the non-sampled companies.
3. The new stratification looks promising. It appears to be better for residential, and not much different for industrial and commercial.

Residential

1. Including IOUs with non-IOUs does not help.
2. Beta=1 is not as good.

Commercial

1. Including IOUs with non-IOUs does help
2. Beta=1 is pretty good. It is better than many alternatives but not necessarily better than two stage least squares.

Industrial

1. Including IOUs with non-IOUs helps with the old stratification but not the new.
2. Beta=1 is best.

Questions for the Committee

1. Do you have alternative suggestions for the summary statistics to use for assessing alternatives?

2. Do you have any suggestions for follow-on studies, particularly the simulation study?

3. Any suggestions for how to best package the results of this study to convince managers to implement changes.

Residential Revenue

Current stratification -- IOUs not used in estimation

experiment	Percent Ave Est Error	Percent RMSE Est Error	Percent CV - Est Error	Percent Error in NS	Percent Error in US total
Beta=1	0.00	0.00	12.76	1.27	0.26
Gamma=.5	0.00	0.00	12.10	1.53	0.32
Gamma=.8	0.00	0.02	66.77	11.30	2.34
Gamma=.5 with outlier	0.00	0.00	10.55	-0.26	-0.05
Gamma=.8 with outlier	0.00	0.00	10.69	0.43	0.09
Two stage	0.00	0.00	11.94	1.22	0.25
Two stage with outlier	0.00	0.00	9.81	0.13	0.03

Current stratification -- IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
beta=1	0.00	0.00	12.76	1.27	0.26
Gamma=.5	0.00	0.00	11.51	1.72	0.36
Gamma=.8	0.00	0.02	45.03	8.16	1.69
Gamma=.5 with outlier	0.00	0.00	11.24	1.19	0.25
Gamma=.8 with outlier	0.00	0.00	11.62	1.29	0.27
Two stage	0.00	0.00	11.45	1.33	0.28
Two stage with outlier	0.00	0.00	10.92	1.01	0.21

New stratification -- IOUs not used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
beta=1	0.00	0.00	12.76	1.27	0.26
Gamma=.5	0.00	0.00	9.88	0.12	0.02
Gamma=.8	0.00	0.02	53.98	5.98	1.24
Gamma=.5 with outlier	0.00	0.00	7.97	-0.83	-0.17
Gamma=.8 with outlier	0.00	0.00	7.91	-0.31	-0.06
Two stage	0.00	0.00	10.69	-0.10	-0.02
Two stage with outlier	0.00	0.00	8.65	-0.87	-0.18

New stratification -- IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
beta=1	0.00	0.00	12.76	1.27	0.26
Gamma=.5	0.00	0.00	9.92	1.58	0.33
Gamma=.8	0.00	0.02	44.42	6.13	1.27
Gamma=.5 with outlier	0.00	0.00	8.60	0.91	0.19
Gamma=.8 with outlier	0.00	0.00	8.57	0.79	0.16
Two stage	0.00	0.00	10.85	1.01	0.21
Two stage with outlier	0.00	0.00	8.86	0.82	0.17

Commercial generation

Current stratification -- IOUs not used in estimation

experiment	Percent Ave Est Error	Percent RMSE Est Error	Percent CV - Est Error	Percent Error in NS	Percent Error in US total
Beta=1	0.00	0.03	67.05	10.21	1.46
Gamma=.5	0.00	0.03	83.94	8.54	1.22
Gamma=.8	0.01	0.05	121.32	14.30	2.05
Gamma=.5 with outlier	0.00	0.03	66.87	6.23	0.89
Gamma=.8 with outlier	0.00	0.03	88.39	9.39	1.35
Two stage	0.00	0.03	67.52	1.69	0.24
Two stage with outlier	0.00	0.02	64.44	5.81	0.83

Current stratification -- IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Beta=1	0.00	0.03	67.05	10.21	1.46
Gamma=.5	0.00	0.03	73.29	8.75	1.25
Gamma=.8	0.01	0.04	108.01	15.50	2.22
Gamma=.5 with outlier	0.00	0.02	57.35	4.90	0.70
Gamma=.8 with outlier	0.00	0.02	61.69	5.79	0.83
Two stage	0.00	0.02	57.43	1.91	0.27
Two stage with outlier	0.00	0.02	53.17	3.26	0.47

New stratification -- IOUs not used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Beta=1	0.00	0.03	67.05	10.21	1.46
Gamma=.5	0.00	0.03	80.20	7.25	1.04
Gamma=.8	0.00	0.04	113.97	11.16	1.60
Gamma=.5 with outlier	0.00	0.03	71.31	5.00	0.72
Gamma=.8 with outlier	0.00	0.03	77.38	7.25	1.04
Two stage	0.00	0.03	71.75	-2.18	-0.31
Two stage with outlier	0.00	0.02	56.55	3.44	0.49

New stratification -- IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Beta=1	0.00	0.03	67.05	10.21	1.46
Gamma=.5	0.00	0.03	75.55	9.98	1.43
Gamma=.8	0.01	0.04	104.88	14.15	2.03
Gamma=.5 with outlier	0.00	0.02	61.25	6.04	0.87
Gamma=.8 with outlier	0.00	0.03	67.38	6.16	0.88
Two stage	0.00	0.02	56.46	0.91	0.13

Two stage with outlier	0.00	0.02	56.68	4.61	0.66

Industrial generation

Current stratification -- IOUs not used in estimation

experiment	Percent Ave Est Error	Percent RMSE Est Error	Percent CV - Est Error	Percent Error in NS	Percent Error in US total
Beta=1	0.00	0.03	56.03	1.78	0.27
Gamma=.5	0.00	0.04	61.25	6.50	1.00
Gamma=.8	0.01	0.06	102.53	17.52	2.70
Gamma=.5 with outlier	0.00	0.03	56.95	1.61	0.25
Gamma=.8 with outlier	0.00	0.03	57.40	1.36	0.21
Two stage	0.00	0.03	57.46	0.92	0.14
Two stage with outlier	0.00	0.03	59.20	-0.30	-0.05

Current stratification -- IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Beta=1	0.00	0.03	56.03	1.78	0.27
Gamma=.5	0.00	0.03	57.05	5.19	0.80
Gamma=.8	0.01	0.04	76.93	12.16	1.88
Gamma=.5 with outlier	0.00	0.03	56.29	3.46	0.53
Gamma=.8 with outlier	0.00	0.03	56.51	3.38	0.52
Two stage	0.00	0.03	55.91	2.08	0.32
Two stage with outlier	0.00	0.03	56.58	0.86	0.13

new stratification -- IOUs not used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Beta=1	0.00	0.03	56.03	1.78	0.27
Gamma=.5	0.00	0.03	60.02	5.52	0.85
Gamma=.8	0.01	0.04	77.15	13.16	2.03
Gamma=.5 with outlier	0.00	0.03	56.14	1.78	0.27
Gamma=.8 with outlier	0.00	0.03	56.05	1.18	0.18
Two stage	0.00	0.03	57.03	0.60	0.09
Two stage with outlier	0.00	0.03	57.55	0.72	0.11

new stratification -- IOUs used in estimation

experiment	Ave Est Error	RMSE Est Error	CV - Est Error	Error in NS	Error in US total
Beta=1	0.00	0.03	56.03	1.78	0.27
Gamma=.5	0.00	0.03	59.14	5.98	0.92
Gamma=.8	0.01	0.05	79.97	17.45	2.69
Gamma=.5 with outlier	0.00	0.03	57.36	3.53	0.54
Gamma=.8 with outlier	0.00	0.03	57.80	4.18	0.64
Two stage	0.00	0.03	60.21	1.63	0.25
Two stage with outlier	0.00	0.03	56.94	3.06	0.47

Stratification

Stratification is intended to describe the unique seasonality associated with each estimation group. The data used for this part of the project is the monthly reported data by company by state, divided by the annual average from that company/state. For each company this provides a series that varies around 1. The peaks in the monthly series occur roughly in February and August.

We tried applying cluster analysis to group states. Using the data for February and August. This was done separately for 2002 and 2003. results are not entirely consistent, but provided a good starting point.

I did a rough eyeball average to come up with the range of numbers for the two years by state, looked at the cluster analysis and came up with the following groups.

Remember this is based on data, not common sense

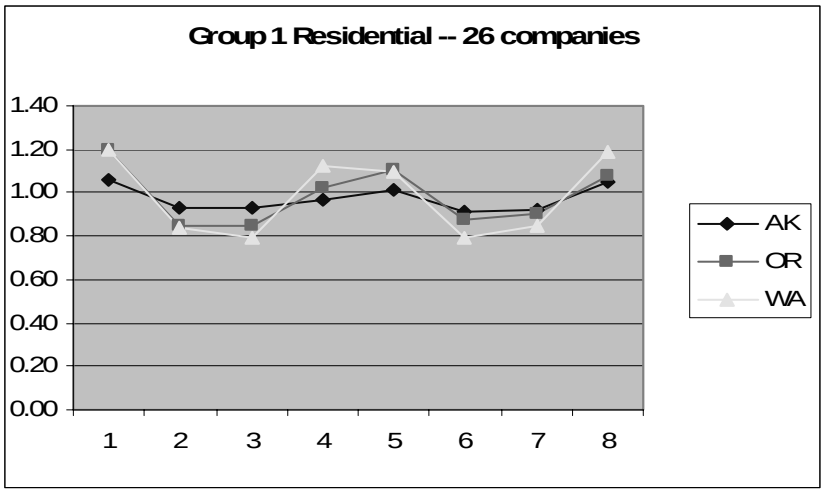
1. AK, WA, OR (1.2, .8)
2. ID, MT, WY, ND (1.15, 1)
3. SD, CO, NM, TX, (.95, 1.2)
4. NV, UT, AZ, OK, KS (.8, 1.6)
5. MN, WI, MI (.95, 1.3)
6. MO, IA. IL, NJ (1, 1.5)
7. AR, MS, IN, KY, AL, LA, GA (.9, 1.3)
8. NC, VA, DE, SC, MD, DC, OH, WV (1.1, 1.25)
9. ME, NH, CT, MA, RI, VT, PA, NY (1.0, 1.2)
10. CA, FL (.85, 1.15)

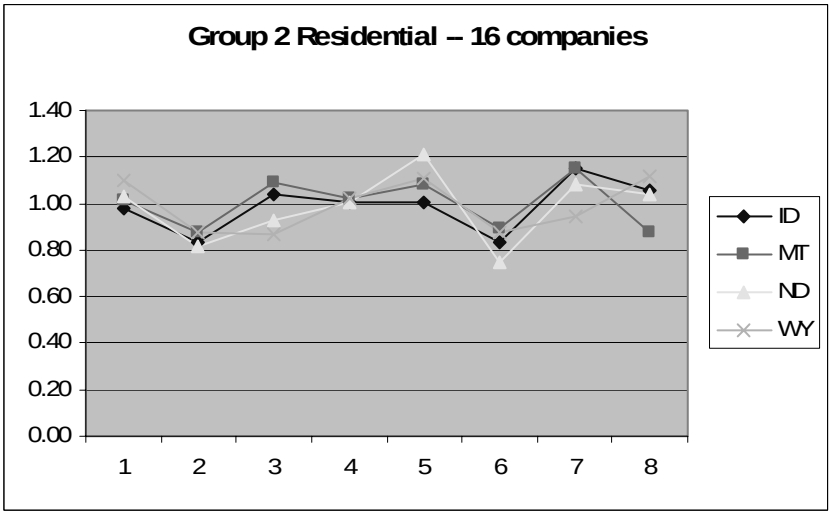
TN is just plain weird. It has seasonality more like the pacific northwest than any nearby state (1.19, .94) 2002 and (1.59, .96) for 2003.

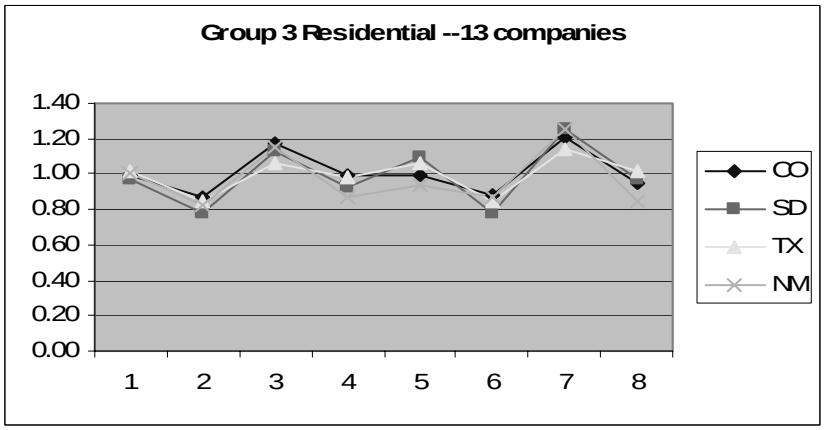
Most likely groups 5, 6 and 10 have too few respondents.

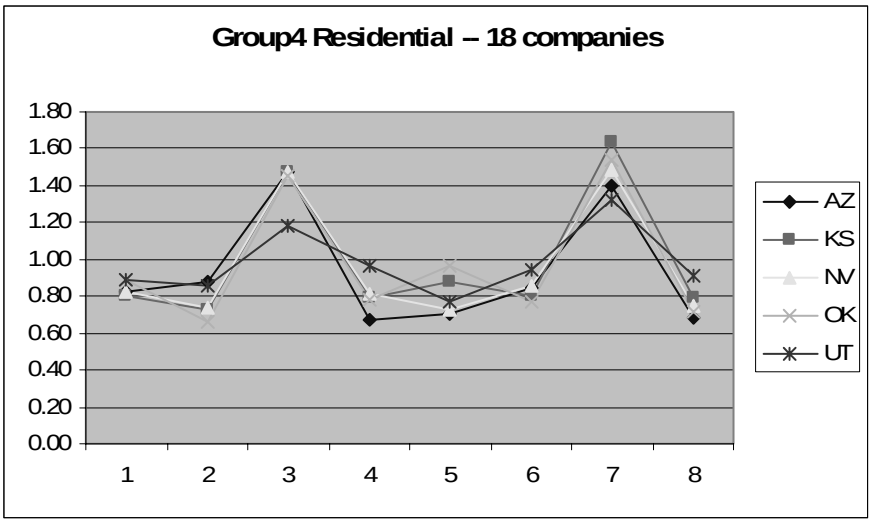
We need to try a more systematic (repeatable) approach for doing this. My general approach was to sort by February then classify into (say) 3 groups – then sort by February-August to determine the second grouping.

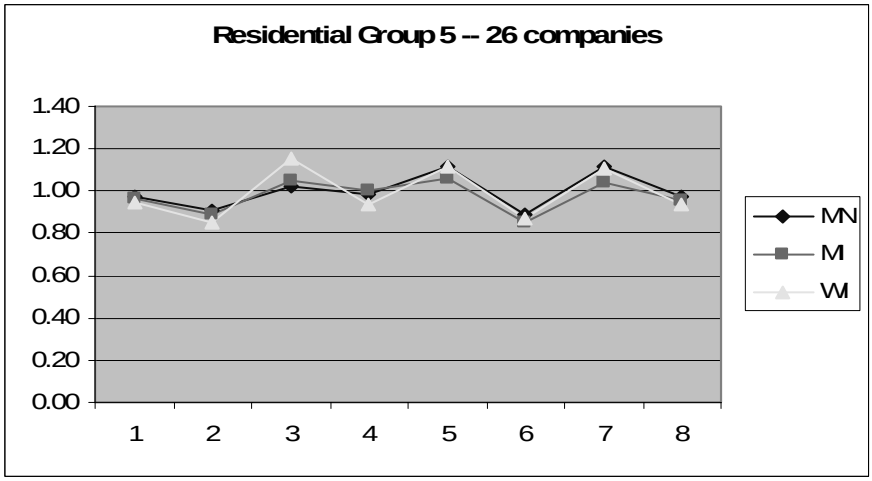
There are interesting patterns. Group 1 is essentially a sine wave with one peak per year in the winter. The other regions have two peaks in the year, one in the winter and one in the summer. However the magnitude of the peaks varies a lot.

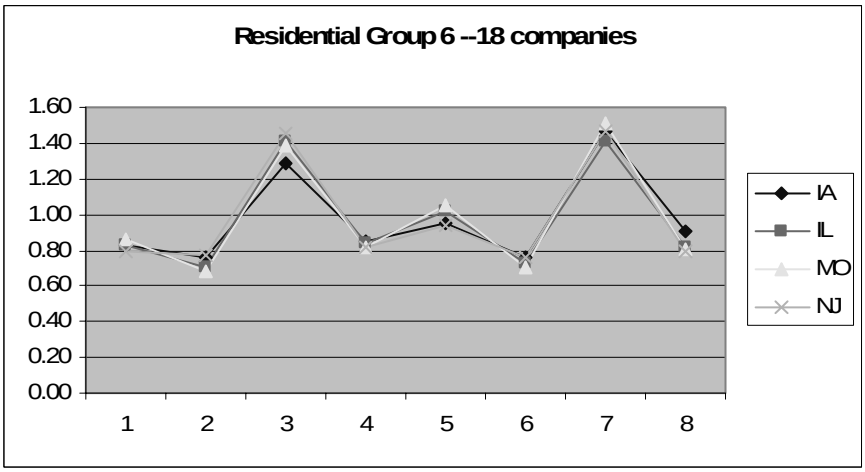

Group 1 Residential -- 26 companies
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1 2 3 4 5 6 7 8
AK
OR
WA

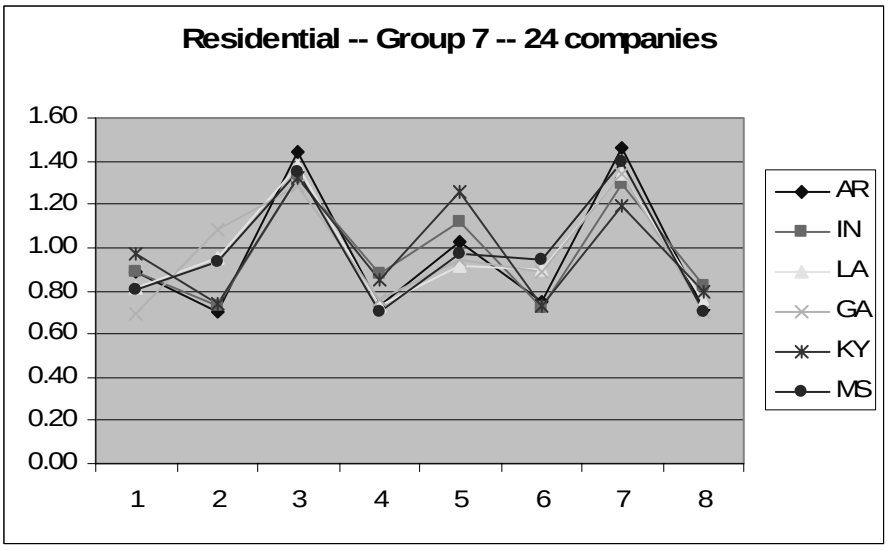

Group 2 Residential – 16 companies
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1 2 3 4 5 6 7 8
ID
MT
ND
WY

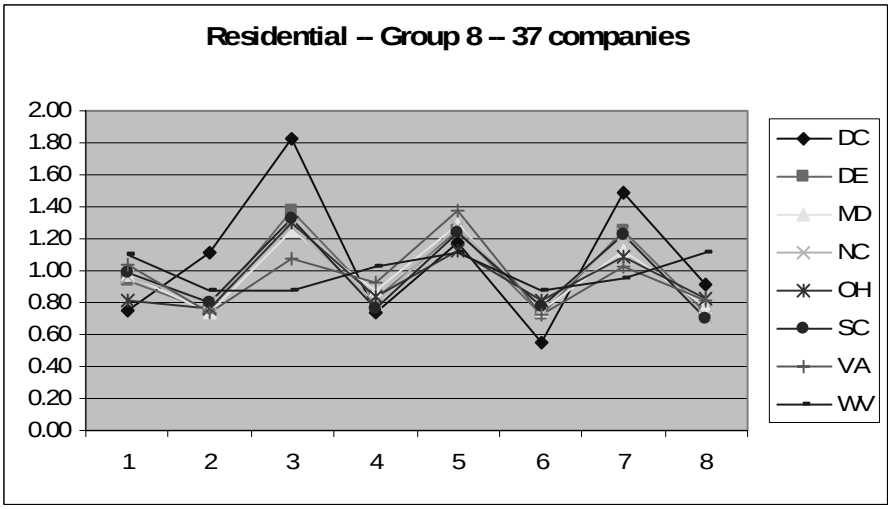

Group 3 Residential --13 companies
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1 2 3 4 5 6 7 8
CO
SD
TX
NM

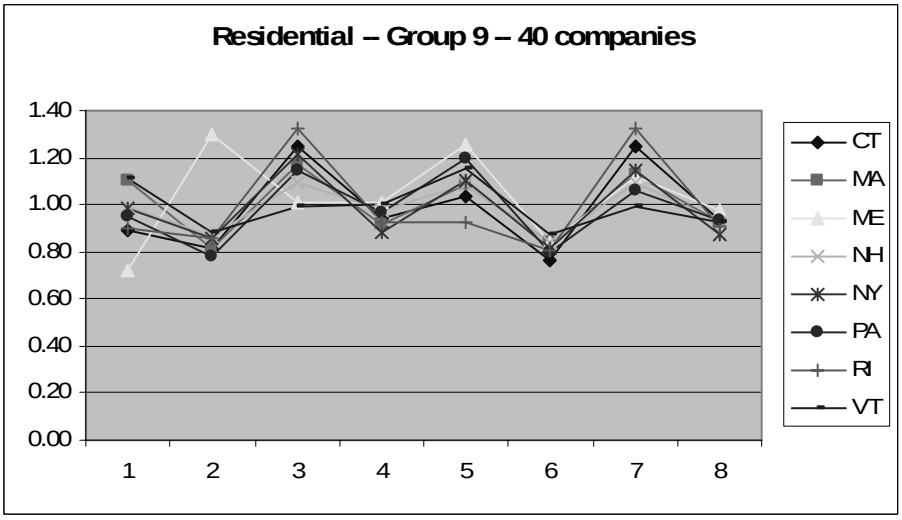

Group4 Residential – 18 companies
1.80
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1
2
3
4
5
6
7
8
AZ
KS
NV
OK
UT

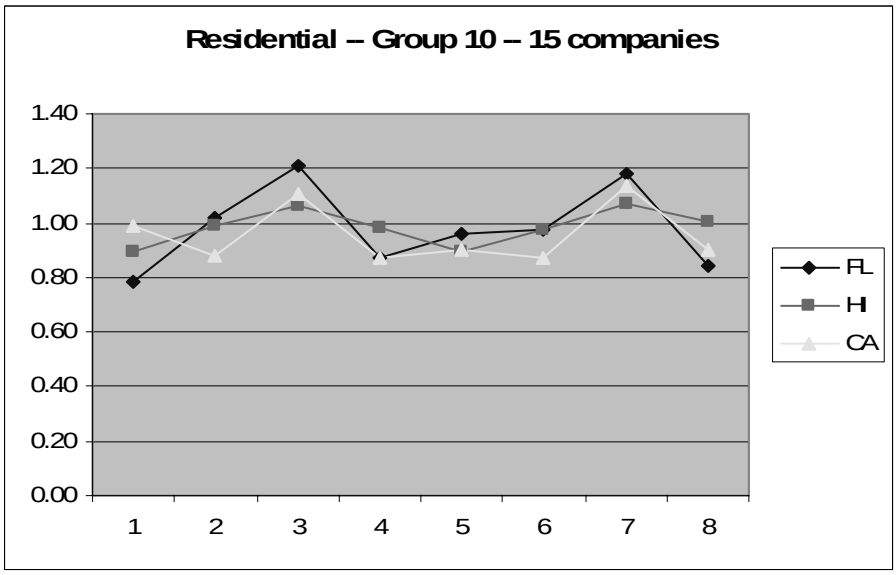

Residential Group 5 -- 26 companies
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1
2
3
4
5
6
7
8
MN
MI
WI

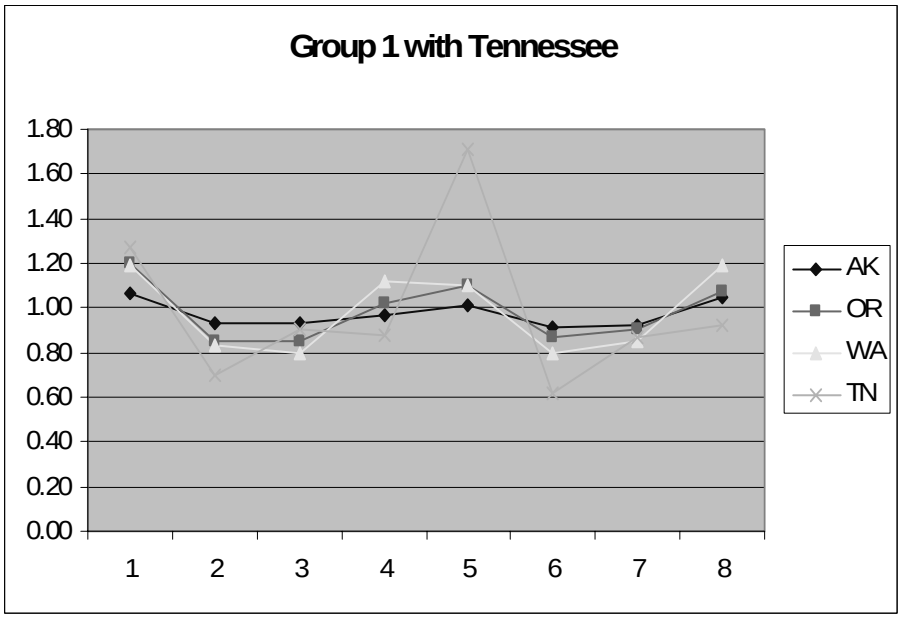

Residential Group 6 --18 companies
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1
2
3
4
5
6
7
8
IA
IL
MO
NJ


Residential -- Group 7 -- 24 companies
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1
2
3
4
5
6
7
8
AR
IN
LA
GA
KY
MS


Residential – Group 8 – 37 companies
2.00
1.80
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1
2
3
4
5
6
7
8
DC
DE
MD
NC
OH
SC
VA
WV


Residential – Group 9 – 40 companies
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1
2
3
4
5
6
7
8
CT
MA
ME
NH
NY
PA
RI
VT





We can assess these groupings by looking at the AIC. I used a simple model, normal distribution with different mean and variance for each quarter. All companies in a state (or group) assumed to follow the same model – use to estimate parameters.

Can compare AIC of group to sum of AICs over states in a group (lower is better)
Can use AIC to classify state to existing groups.

We may do some analysis of other data to see if there are ways to explain some of these groupings. Weather is clearly not the only factor. Others include percent of customers who use electricity (or other fuels) for heating; incentives that may have been offered to customers to convert them to using electricity.

Do you have other thoughts and/or ideas about stratification?